International Journal of Research in Marketing 33 (2016) 944–960

Contents lists available at ScienceDirect

IJRM
International Journal of Research in Marketing
journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

The calibrated sigma method: An efficient remedy for between-group differences in response category use on Likert scales

Bert Weijters [a,1], Hans Baumgartner [b,2], Maggie Geuens [c,d,3]

[a] Department of Personnel Management, Work and Organizational Psychology, Ghent University, Dunantlaan 2, B-9000 Ghent, Belgium
[b] The Pennsylvania State University, Department of Marketing, 482 Business Building, University Park, PA 16802, United States
[c] Ghent University, Tweekerkenstraat 2, B-9000 Ghent, Belgium
[d] Vlerick Business School, Reep 1, B-9000 Ghent, Belgium

ARTICLE INFO

Article history:
First received on May 7, 2015 and was under review for 10 months
Available online 7 June 2016

Area Editor: Stefano Puntoni

Keywords:
Response bias
Language differences
Survey methods
Likert items

ABSTRACT

The authors propose a procedure, labeled the calibrated sigma method, which is designed to correct for between-group differences in endorsement likelihood of response categories that are unrelated to the content of the items. The method is especially useful in cross-cultural research where group differences may reflect variation in scale usage rather than substantive differences. However, the procedure is also relevant in other situations, for example, when different data collection modes or different experimental manipulations affect respondents' perception of the meaning of the scale labels. The calibrated sigma method uses information derived from heterogeneous control items (calibration items) to reweight the responses to substantive items in a group-specific way. The advantages of the calibrated sigma method are that it avoids the arbitrariness in the assignment of particular numerical values to response categories; that it is compatible with the linear model, which is used by most marketing researchers; and that it does not require the use of complex nonlinear models involving the estimation of many additional measurement model parameters. The authors validate the calibrated sigma method on a simulated cross-linguistic data set pertaining to 12 different languages; an empirical data set collected from respondents of the same nationality but from two different language groups; and an experimental data set consisting of responses to two different response scale formats. The findings demonstrate that the proposed procedure controls for artefactual scale use differences across groups but does not eliminate substantive differences. It is particularly efficient for marketing research agencies, panel providers and other marketing researchers who analyze surveys involving multiple language groups, different scale formats, multiple modes of data collection, or different manipulations affecting the meaning of the response category labels.

1. Introduction

When researchers want to compare scores on variables of interest across groups or conditions, scale usage heterogeneity is an important source of concern. The term scale usage heterogeneity (also called differential scale usage) refers to systematic differences in how respondents in different groups use the response scale, which are unrelated to substantive differences on the

E-mail addresses: bert.weijters@ugent.be (B. Weijters), hansbaumgartner@psu.edu (H. Baumgartner), Maggie.geuens@UGent.be (M. Geuens).
[1] Tel.: +32 9264 62 96; fax: +32 9264 64 94.
[2] Tel.: +1,814,863 3559.
[3] Tel.: +32 9264 35 21; fax: +32 9264 42 79.

http://dx.doi.org/10.1016/j.ijresmar.2016.05.003
0167-8116/

variables studied. Scale usage heterogeneity is problematic because it may lead to artificial differences between groups or mask true differences.

Scale usage differences are often conceptualized as individual differences that should be assessed and controlled at the respondent level (Baumgartner & Steenkamp, 2001; Fischer, 2004; Rossi, Gilula, & Allenby, 2001). However, in certain situations scale usage heterogeneity may occur primarily at the group level, in which case it is more appropriate to model differential scale usage at the group level. For example, when Likert-type rating scales anchored by labels such as 'strongly (dis)agree' or 'completely (dis)agree' are used in different languages, the meaning of the response category labels may subtly but systematically vary across languages, which can lead to differences in scale usage at the group level (Skevington & Tucker, 1999; Smith, Mohler, Harkness, & Onodera, 2005; Szabo, Orley, & Saxena, 1997; Weijters, Geuens, & Baumgartner, 2013). Similarly, data collection modes or experimental manipulations may affect the perceived meaning of the category labels and thus induce scale usage heterogeneity (Jordan, Marcus, & Reeder, 1980; Weijters, Schillewaert, & Geuens, 2008). In these cases, different response distributions across groups are not due to item content, but occur because of the non-equivalence of response category meanings.

In an attempt to remedy this potential bias, we introduce a procedure labeled the calibrated sigma method, which is designed to eliminate the non-comparability of responses across groups (e.g., cultures, languages, modes of data collection, experimental conditions) at the group, rather than individual, level. Instead of assigning the same consecutive integers to the scale positions in all groups (e.g., in the case of a 5-point scale, 'strongly disagree' is usually coded as 1, 'disagree' as 2, 'neither agree nor disagree' as 3, 'agree' as 4, and 'strongly agree' as 5), the response categories are converted to numerical values in a group-specific way. Specifically, the numbers assigned to the response categories are based on the distribution of responses to an independent and heterogeneous set of control items, which serve no purpose other than assessing the content-free endorsement frequencies of the response categories in different groups (i.e., these calibration items are not used for substantive purposes). Thus, instead of arbitrarily assuming an equal-interval scale, the scale scores are chosen based on how the different groups respond to a set of content-free items, or at least items that share no obvious common content. For instance, 'strongly agree' might be coded as 5 in English, whereas 'tout à fait d'accord' is coded as 4.5 in French, corresponding to the different endorsement rates of the fifth option in response to the control items across the two languages.

After presenting an overview of previous approaches to dealing with scale usage differences at the individual level and a detailed description of the proposed procedure, we present three complementary studies in the current paper. In the first study, we use a simulated data set to illustrate how the proposed calibrated sigma method works, based on a comparison of traditionally coded and sigma coded responses simulated for twelve different languages, and we show how the new procedure can yield more valid results than the conventional procedure. Specifically, in contrast to the traditional procedure, the new procedure does not indicate artificial group differences in case there are none while it does not wash out genuine differences. This study also demonstrates that testing for measurement invariance across groups will not identify scale usage differences when the bias is uniform across items. In the second study, we apply the calibrated sigma method to an empirical data set of respondents who share the same nationality (Belgian) but use different languages (Dutch and French), and we demonstrate that the new procedure leads to conclusions that differ from the conventional method but are consistent with the results of an analysis that corrects for response styles at the individual level. In particular, while the conventional method suggests that there might be a significant difference in the construct of interest between Dutch- and French-speaking respondents, the calibrated sigma method and the individual-level response style correction method both indicate that this difference is most likely caused by scale usage differences. In the third study, we illustrate the potential use of the proposed method in an experimental context in which survey responses are obtained with two alternatively labeled response scale formats to which respondents are randomly assigned. We demonstrate that calibrated sigma coding outperforms traditional coding and leads to results that are comparable to those obtained with more elaborate and involved individual-level response style correction methods.

2. Literature review

It is well-known in the survey literature that observed scores on variables of interest contain not only substantive but also non-content-related sources of variation. The term common method bias is often used to refer to the general problem of non-random variance in measures that is independent of content (Podsakoff, MacKenzie, & Podsakoff, 2012). In this paper we are specifically concerned with systematic differences in how respondents use the response scale (scale usage heterogeneity). Usually, differential scale usage is conceptualized as a respondent-specific phenomenon, such that different respondents vary in their preference for certain scale positions. Two broad approaches to controlling for individual-level scale usage heterogeneity can be distinguished.

In the first approach, the items measuring the substantive constructs of interest are used to assess and correct for differences in scale usage, and the sources of differential scale usage are not identified in detail. A popular method exemplifying this approach is to standardize (or at least mean-center) the data within respondents. That is, a person's responses to the substantive items are converted into z-scores by subtracting from each response the respondent's mean response across all items and dividing by the standard deviation of the respondent's ratings (Fischer, 2004). This method acknowledges that there may be systematic differences in the level and spread of people's responses across items, but otherwise the sources of scale usage heterogeneity are left unexplored. Although the method is simple, there are three problems. First, the procedure assumes that the raw data contain interval information, even though ratings probably only yield ordinal data. Second, the within-person estimates of scale usage (means and standard deviations) may not be very reliable, particularly if they are based on few responses. Third and most importantly, the respondent-specific means and standard deviations are supposed to be "pure" measures of scale usage, but since they are based on the same items for which substantive analyses are to be conducted, it is likely that scale usage will be confounded with content.